### Author Names & Affiliations

- Caroline Laplante - Department of Molecular Biomedical Sciences, College of Veterinary Medicine, North Carolina State University
- Michela Becchi - Department of Electrical and Computer Engineering, North Carolina State University
- Cranos Williams - Department of Electrical and Computer Engineering, North Carolina State University
- Mary Elting - Cell and Tissue Biology Department, UCSF. Starting August 2017: Physics Department, North Carolina State University

### Contact Email Address (for NSF use only)

(Hidden)

### Research Domain, discipline, and sub-discipline

Cellular Biology, Quantitative Microscopy, SMLM

### Title of Submission

A high-performance computing workflow for quantitative microscopy data

### **Abstract** (maximum ~200 words).

With recent progress in single molecule localization microscopy (SMLM), the field of quantitative microscopy is growing rapidly. Novel measurements of molecular concentrations and localizations are beginning to reveal how biomolecules move, organize, and adapt within live cells and tissues. Interdisciplinary collaborative efforts among biologists, mathematicians, physicists, and engineers synergize to drive innovative, quantitative approaches to understanding biological processes. Quantitative results fuel mathematical models that generate new experimental hypotheses, broadening our understanding and perspective of biological systems.

The time required to produce meaningful quantitative results after acquisition of raw microscopy data hinders the efficiency of the process. Large raw datasets are cumbersome to transfer from acquisition platforms to analysis computers. These datasets are even more difficult to share among collaborative groups especially since current image processing strategies lack streamlining and unifying standardization.

An efficient pipeline that allows for the processing, analysis, visualization, and sharing of large datasets must be established for the success of such interdisciplinary, collaborative groups. Advanced computational tools have the potential to effectively bridge the gap between biological experimental approaches and meaningful quantitative analyses, improving the workflow of experimental designs and formulation of novel hypotheses.

### **Question 1** Research Challenge(s) (maximum ~1200 words): Describe current or emerging science or engineering research challenge(s), providing context in terms of recent research activities and standing questions in the field.

Achieving the ultimate goal of modeling development in silico requires quantification of in vivo dynamics at the tissue, cellular, and molecular levels. An emerging field of biology that takes significant steps toward this goal is quantitative high-speed SMLM to elucidate protein organization in living organisms. SMLM methods such as PALM, FPALM or STORM (Betzig et al., 2006; Hess et al., 2006; Rust et al., 2006), whose demonstration over 10 years ago led to the Nobel Prize in Chemistry in 2014, can resolve biological structures at the nanometer scale (~20 nm lateral resolution), a tenfold improvement in resolution over traditional light microscopy techniques. The fundamental principle of SMLM is the sequential imaging of sparse subsets of fluorophores distributed over thousands of camera frames. The center of each fluorophore on each camera frame is precisely localized, generating a dataset that can be used to reconstruct high-resolution images. This technique has now been adapted to live cell imaging, generating gigabyte-size datasets with thousands of camera frames per minute of acquisition (Huang et al., 2013; Laplante et al., 2016a; Laplante et al., 2016b). Exploiting SMLM to characterize live cells under various conditions would require large-scale acquisition and analysis of many microscopy datasets. These demanding tasks drive the need for automated image-analysis algorithms to correct camera noise, detect emission and precisely locate individual fluorescent molecules taking advantage of their separation in space and time. The lack of such approaches results in the inability of biologists and mathematicians to exploit the full quantitative power of this method.

SMLM has already proven to be a powerful tool in the discoveries of new cellular structures, for example, with the identification of architectural rings within neurons that are blurred together and unrecognizable when neurons are imaged with traditional light microscopy techniques (Xu et al., 2013). Most often SMLM is used to complement a research project with a high-resolution image of a relevant cellular structure. Yet, the tremendous potential for SMLM lies in the quantitative data contained in the precise position of single-molecules in space and time. Quantitative SMLM analyses were recently developed to elucidate the organization of proteins within the contractile cytokinetic apparatus in fission yeast (Laplante et al., 2016b). Few biologists however take advantage of the quantitative power of this method in part because of the complex processing of such large datasets. Exploiting SMLM to characterize live cells under various conditions would require large-scale acquisition and analysis of many microscopy datasets. These demanding tasks drive the need for automated image-analysis algorithms to correct camera noise, detect emission and precisely locate individual fluorescent molecules taking advantage of their separation in space and time. Obstacles that hinder the broad utility of SMLM include dealing with large data files, difficulties in optimizing acquisition parameters a priori, and rapidly evolving SMLM techniques such as the advent of faster more sensitive cameras, the development of new dyes and fluorescent proteins, and the need for acquisitions in three dimensions. The lack of such approaches results in the inability of biologists and mathematicians to exploit the full quantitative power of this method.

In traditional microscopy techniques, a displayed image is visible in real time as the sample is imaged. However, SMLM comes at the cost of substantial processing time to achieve meaningful results due to its gigabyte-size datasets. For example, the analysis of 10,000 frames of ~1.5 GB size on a commodity machine may require 10 minutes or longer to localize all the single molecules. SMLM data are first acquired with no knowledge of quality until the processing of raw data is complete and the data are reconstructed into an image that can be visually evaluated. Such lengthy processing times preclude optimizing sample preparation and imaging conditions to obtain ideal single molecule localization and eliminate random background noise from cellular cytoplasm to increase calculated precision.

As more fluorescent proteins and membrane permeable dyes are developed for SMLM, the task will become more complex. Dual-color imaging requires the concomitant fitting of fluorophores of different photo-physical properties within datasets and performing sensitivity controls and optimization for each population of fluorescent molecules. Furthermore, SMLM techniques can provide information about fluorophores in all three dimensions (3D). The fitting of single molecules in 3D, however, taxes the analysis even more and often results in loss of data due to the rejection of a large number of single molecules.

Other types of microscopy utilize single molecule fitting and tracking both in vivo and in vitro. In particular, fluorescence speckle microscopy (FSM) includes many of the same analysis challenges as SMLM and would greatly benefit from an advanced image-processing pipeline. In FSM, samples express sparse populations of proteins tagged with fluorophores (Waterman-Storer et al., 1998) that need to be identified and localized before meaningful quantitative data can be extracted. The recent development of SunTag, a novel protein tag that amplifies the signal-to-noise ratio, extends this approach to the single molecule level (Tanenbaum et al., 2014). As in SMLM, fluorescent molecules must be localized in each camera frame, but with FSM, subsequent frames are correlated to identify the trajectories of individual molecules. Effective algorithms to more efficiently and accurately extract these data would vastly increase the speed and volume of information measured through this technique.

These challenges can be overcome by developing a flexible and configurable high-performance computing pipeline that, while greatly reducing the time from data acquisition to the availability of results, would allow the user to easily configure the image visualization and analysis.

**Question 2** Cyberinfrastructure Needed to Address the Research Challenge(s) (maximum ~1200 words): Describe any limitations or absence of existing cyberinfrastructure, and/or specific technical advancements in cyberinfrastructure (e.g. advanced computing, data infrastructure, software infrastructure, applications, networking, cybersecurity), that must be addressed to accomplish the identified research challenge(s).

There is a great need in the SMLM field and the broader realm of microscopy imaging for a cyber ecosystem combining a workflow for the identification of the fluorophores and a shared and searchable data storage system for distributing data among collaborators. Currently, numerous researchers are generating vast amounts of SMLM data, but image processing and data transfer create significant bottlenecks in analyzing and interpreting such data. Research groups who acquire SMLM data but do not have the expertise to perform such complex quantitative analysis would greatly benefit from accessing analysis algorithms remotely through an intuitive and easy-to-use interface. A computing workflow including rapid data transfer from instrument to analysis platforms, parallel processing of complex large-scale data and formatting of quantitative data for mathematical modeling would shorten the elapsed time between the experimental design to the mathematical modeling process.

In order to achieve fast and efficient quantitative results from raw SMLM data, each step of the analysis - from the localization of each single molecule to the calculation of localization uncertainty - needs to be accelerated. The use of parallel processing techniques and platforms in the single molecule localization step would improve the overall efficiency of the process. This would produce nearly simultaneous reconstruction of high quality still images and time-lapse videos for visual qualitative evaluation of the data during acquisition. These same techniques could also be used to analyze FSM data, with the additional task of correlating locations between subsequent frames following the initial localization. This effective processing would have the same benefits for all microscopy data dealing with either single molecules or protein particles – improved accuracy, speed, and efficiency in quantitative data analysis.

Graphics Processing Units (GPUs), widely used to accelerate applications from computational biology, chemistry, physics, numerical analytics, weather prediction, and data mining (among other domains), are naturally suited to accelerate localization of molecules within a single image. Besides data localization within images, SMLM requires high-speed analysis across multiple frames. A computing pipeline that, leveraging different computing platforms (e.g., CPUs and GPUs in a cluster setting), combines fine- and coarse-grained parallelization within and across images, is required to perform high-speed data processing on large datasets. Further, in order to enable different analyses, this pipeline should be configurable and extensible. Finally, to support multiple users, the pipeline should include seamless resource allocation and scheduling mechanisms. While image processing acceleration and resource allocation have been extensively studied (and are still subjects of research) within the computing community, a platform leveraging these techniques for the analysis of quantitative microscopy data is missing.

The storage of large imaging datasets is a common challenge among biologists. The data are often temporarily stored on acquisition computers and then transferred onto external hard drives. With time, valuable, hard-earned data become misplaced or forgotten and biologists repeat experiments that were already performed wasting both time and resources. Organizing and storing data in a centralized system would prevent repeating experiments unnecessarily, allow other collaborating groups to search and utilize previously acquired datasets and expand the amount of results obtained by re-using data to answer new questions. This would require a standardized record-keeping and classification of image metadata, a searchable image library and the possibility to safely download raw image data remotely. When implemented on a small scale, such a system would benefit members of a single group or department in keeping accurate records of their acquired data. The benefits would increase when used among multiple groups of collaborators spread across separate departments or universities. A greater vision would be to build a data analysis and storage facility that could be accessed from groups across the globe. The proper functioning of such a system may require the data to be tracked as they are downloaded by other groups for new analyses and publication. Collaborators within universities and eventually across the globe could share data using this system. The Protein DataBase (PDB) is a similar system used for the deposition and the sharing of protein crystal structure data. The research and education computer network consortium Internet2 could be leveraged to allow high-speed data transfer and access to the system from many institutions within the country.

**Question 3** Other considerations (maximum ~1200 words, optional): Any other relevant aspects, such as organization, process, learning and workforce development, access, and sustainability, that need to be addressed; or any other issues that NSF should consider.

Implementing such a cyberinfrastructure would require an interdisciplinary effort of biologists, engineers, programmers and mathematicians to develop analysis methods pertinent to imaging data and improve the computing aspects of the analysis of biological samples. It would also require departmental Information Technology (IT) specialists to establish the infrastructure onto a sharable and accessible platform such as the cloud or Internet2 and to enable fast data transfer from the instrumentation platform located physically in a laboratory to this centralized data analysis platform.

The collaborative groups participating in the elaboration of such a pipeline would provide a prime environment for cross-disciplinary training experience for specialists in each field. We note that the National Research Council’s Committee on Undergraduate Biology Education (BIO2010, 2003) has argued that there is an urgent need to expand biology education to include introductions to computing (including high-performance computing) and mathematical modeling, as computing and mathematics are an essential part of modern biology (Bialek and Botstein, 2004; Cohen, 2004; Pevzner and Shamir, 2009; Robeva and Laubenbacher, 2009). Biologists involved in this high-performance computing workflow for quantitative microscopy data would become well versed in efficient quantitative analysis and in IT applications while IT specialists and programmers would learn about the demands of academics. This large-scale endeavor would be divided into a multitude of projects pertaining to the development of the cyberinfrastructure or of new computational and analysis tools for trainees ranging from undergraduate students to postdoctoral fellows. The long-term vision of such a program is to establish a shared imaging database for both raw data deposition and image analysis tools, and to promote collaboration and cross-pollination between different research communities.

References

Betzig, E., Patterson, G.H., Sougrat, R., Lindwasser, O.W., Olenych, S., Bonifacino, J.S., Davidson, M.W., Lippincott-Schwartz, J., and Hess, H.F. (2006). Imaging intracellular fluorescent proteins at nanometer resolution. Science 313, 1642-1645.

Bialek, W., and Botstein, D. (2004). Introductory science and mathematics education for 21st-Century biologists. Science 303, 788-790.

BIO2010, N. (2003). Transforming Undergraduate Education for Future Research Biologists (Washington, DC: National Academies Press).

Cohen, J.E. (2004). Mathematics is biology's next microscope, only better; biology is mathematics' next physics, only better. PLoS Biol 2, e439.

Hess, S.T., Girirajan, T.P., and Mason, M.D. (2006). Ultra-high resolution imaging by fluorescence photoactivation localization microscopy. Biophys J 91, 4258-4272.

Huang, F., Hartwich, T.M., Rivera-Molina, F.E., Lin, Y., Duim, W.C., Long, J.J., Uchil, P.D., Myers, J.R., Baird, M.A., Mothes, W., et al. (2013). Video-rate nanoscopy using sCMOS camera-specific single-molecule localization algorithms. Nat Methods 10, 653-658.

Laplante, C., Huang, F., Bewersdorf, J., and Pollard, T.D. (2016a). High-Speed Super-Resolution Imaging of Live Fission Yeast Cells. Methods Mol Biol 1369, 45-57.

Laplante, C., Huang, F., Tebbs, I.R., Bewersdorf, J., and Pollard, T.D. (2016b). Molecular organization of cytokinesis nodes and contractile rings by super-resolution fluorescence microscopy of live fission yeast. Proc Natl Acad Sci U S A 113, E5876-E5885.

Pevzner, P., and Shamir, R. (2009). Computing has changed biology--biology education must catch up. Science 325, 541-542.

Robeva, R., and Laubenbacher, R. (2009). Mathematical biology education: beyond calculus. Science 325, 542-543.

Rust, M.J., Bates, M., and Zhuang, X. (2006). Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM). Nat Methods 3, 793-795.

Tanenbaum, M.E., Gilbert, L.A., Qi, L.S., Weissman, J.S., and Vale, R.D. (2014). A protein-tagging system for signal amplification in gene expression and fluorescence imaging. Cell 159, 635-646.

Waterman-Storer, C.M., Desai, A., Bulinski, J.C., and Salmon, E.D. (1998). Fluorescent speckle microscopy, a method to visualize the dynamics of protein assemblies in living cells. Curr Biol 8, 1227-1230.

Xu, K., Zhong, G., and Zhuang, X. (2013). Actin, spectrin, and associated proteins form a periodic cytoskeletal structure in axons. Science 339, 452-456.

### Consent Statement